Lecture 10 - Multiple Regression

Bill Perry

Lecture 09: Review

Covered

- Regression T-Test Anova
- Regression Assumptions
- Model II Regression

Lecture 10: Overview

Multiple Linear Regression model

- Regression parameters
- Analysis of variance
- Null hypotheses
- Explained variance
- Assumptions and diagnostics
- Collinearity
- Interactions
- Dummy variables
- Model selection
- Importance of predictors

Lecture 10: Analyses

What if more than one predictor (X) variable?

- If predictors continuous
- Mix between categorical and continuous
- Can use multiple linear regression

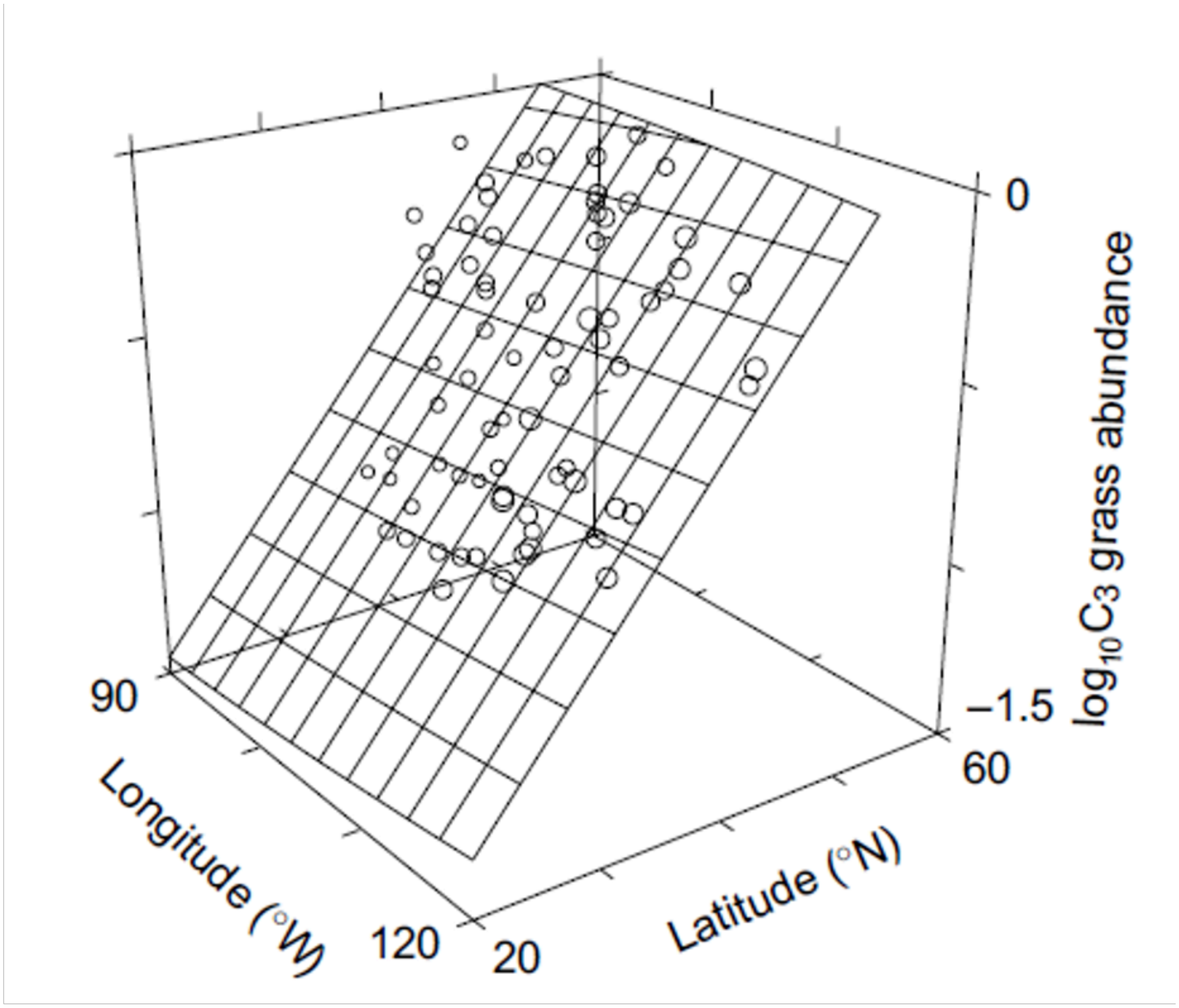
	Independent variable	
Dependent variable	**Continuous**	**Categorical**
Continuous	Regression	ANOVA
Categorical	Logistic regression	Tabular

Lecture 10: Analyses

Abundance of C3 grasses can be modeled as function of

- latitude
- longitude
- both

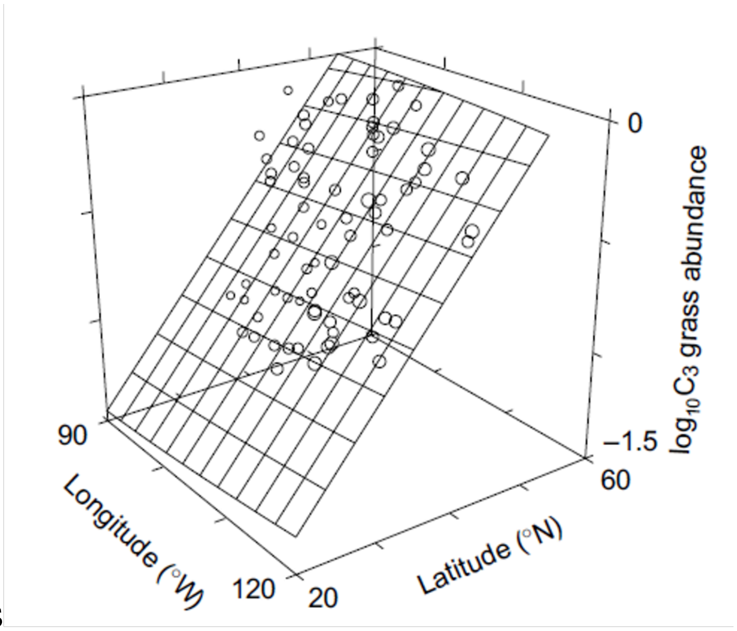
Instead of line, modeled with (hyper)plane



Lecture 10: Analyses

Used in similar way to simple linear regression:

- Describe nature of relationship between Y and X's
- Determine explained / unexplained variation in Y
- Predict new Ys from X
- Find the "best" model



S

Lecture 10: Analyses

Crawley 2012: "Multiple regression models provide some of the most profound challenges faced by the analyst":

- Overfitting
- Parameter proliferation
- Multicollinearity
- Model selection

Lecture 10: Analyses

Multiple Regression:

- Set of i= 1 to n observations
- fixed X-values for p predictor variables (X1, X2…Xp)
- random Y-values:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \epsilon_i$$

- yi: value of Y for ith observation X1 = xi1, X2 = xi2,…, Xp = xip
- β0: population intercept, the mean value of Y when X1 = 0, X2 = 0,…, Xp = 0

Lecture 10: Multiple linear regression model

Multiple Regression:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \epsilon_i$$

- β1: partial population slope, change in Y per unit change in X1 holding other X-vars constant
- β2: partial population slope, change in Y per unit change in X2 holding other X-vars constant
- βp: partial population slope, change in Y per unit change in Xp holding other X-vars constant

Lecture 10: Regression parameters

Multiple Regression:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \epsilon_i$$

- εi: unexplained error - difference bw yi and value predicted by model (ŷi)
- NPP = β0 + β1(lat) + β2 (long) + β3 (soil fertility) + εi

Lecture 10: Regression parameters

Multiple Regression:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \epsilon_i$$

- Estimate multiple regression parameters (intercept, partial slopes) using OLS to fit the regression line
- OLS minimize Σ(yi-ŷi)2, the SS (vertical distance) between observed yi and predicted ŷi for each xij
- ε estimated as residuals: εi = yi-ŷi
- Calculation solves set of simultaneous normal equations with matrix algebra

Lecture 10: Regression parameters

Regression equation can be used for prediction by subbing new values for predictor (X) variables

- Confidence intervals calculated for parameters
- Confidence and prediction intervals depend on number of observations and number of predictors
  - More observations decrease interval width
  - More predictors increase interval width
- Prediction should be restricted to within range of X variables

Lecture 10: Analyses of variance

Variance - SStotal partitioned into SSregression and SSresidual

- SSregression is variance in Y explained by model
- SSresidual is variance not explained by model

Source of variation	SS	df	MS	Interpretation
Regression	$\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2$	p	$\frac{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}{p}$	Difference between predicted observation and mean
Residual	$\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2$	$n-p-1$	$\frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{n-p-1}$	Difference between each observation and predicted
Total	$\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2$	$n-1$		Difference between each observation and mean

Lecture 10: Analyses

SS converted to non-additive MS (SS/df)

- MSresidual: estimate population variance
- MSregression: estimate population variance + variation due to strength of X-Y relationships
- MS do not depend on sample size

Source of variation	SS	df	MS
Regression	$\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2$	p	$\frac{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}{p}$
Residual	$\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2$	$n-p-1$	$\frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{n-p-1}$
Total	$\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2$	$n-1$	

Lecture 10: Hypotheses

Two Hos usually tested in MLR:

- "Basic" Ho: all partial regression slopes equal 0; β1 = β2 = … = βp = 0
- If "basic" Ho true, MSregression and MSresidual estimate variance and their ratio (F-ratio) = 1
- If "basic" Ho false (at least one β ≠ 0) MSregression estimates variance + partial regression slope and their ratio (F-ratio)
- will be > 1 - F-ratio compared to F-distribution for p-value

Lecture 10: Hypotheses

Also: is any specific β = 0 (explanatory role)?

- E.g., does LAT have effect on NPP?
- These Hs tested through model comparison
- Model w 3 predictors X1, X2,X3 (model 1):
- yi= β0 +β1xi1+β2xi2+β3xi3+ εi
- To test Ho that β1 = 0 compare fit of model 1 to model 2:
- yi= β0 +β2xi2+β3xi3+ εi

Lecture 10: Hypotheses

- If SSregression of mod1=mod2, cannot reject Ho β1 = 0
- If SSregression of mod1 > mod2, evidence to reject Ho β1 = 0

- SS for β1 is SSextraβ1 = Full SSregression - Reduced SSregression
- Use partial F-test to test Ho β1 = 0 :

$$F_{w,n-p} = \frac{MS_{Extra}}{FULL\ MS_{Residual}}$$

Can also use t-test (R provides this value)

Lecture 10: Explained variance

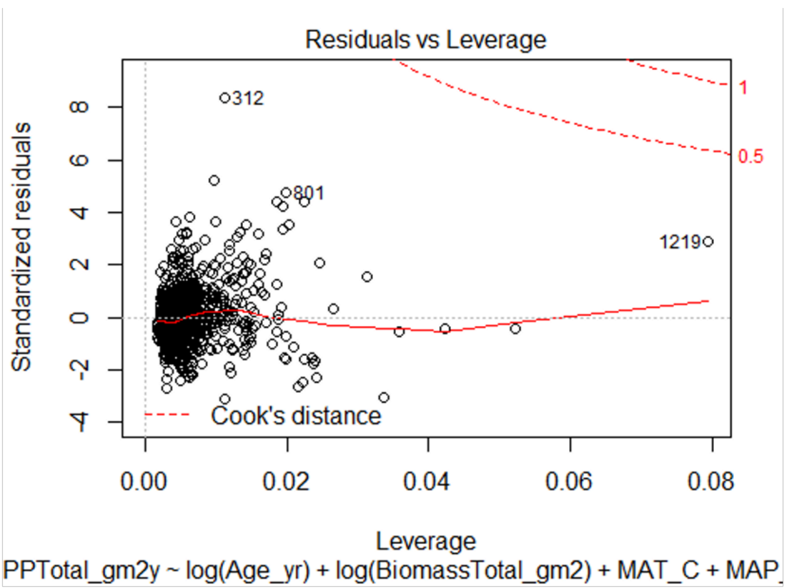
Explained variance (r2) is calculated the same way as for simple regression:

$$r^2 = \frac{SS_{Regression}}{SS_{Total}} = 1 - \frac{SS_{Residual}}{SS_{Total}}$$

- r2 values can not be used to directly compare models
- r2 values will always increase as predictors added
- r2 values with different transformation will differ

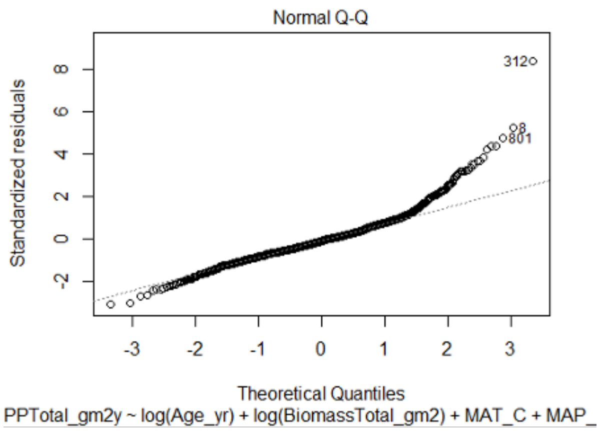
Lecture 10: Assumptions and diagnostics

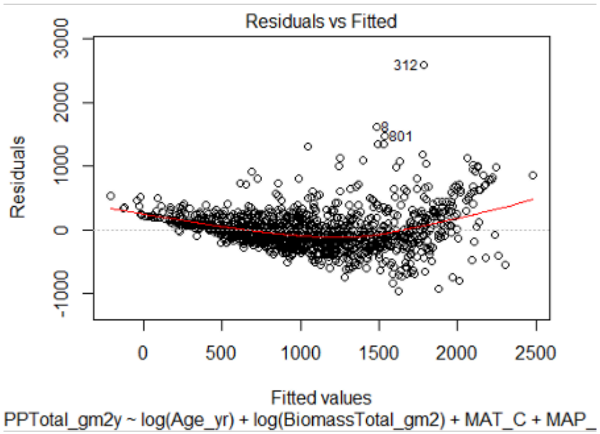
- Assume fixed Xs; unrealistic in most biological settings
- No major (influential) outliers
- Check leverage, influence- Cook's Di



Lecture 10: Assumptions and diagnostics

- Normality, equal variance, independence
- Residual QQ-plots, residuals vs. predicted values plot
- Distribution/variance often corrected by transforming Y





Lecture 10: Assumptions and diagnostics

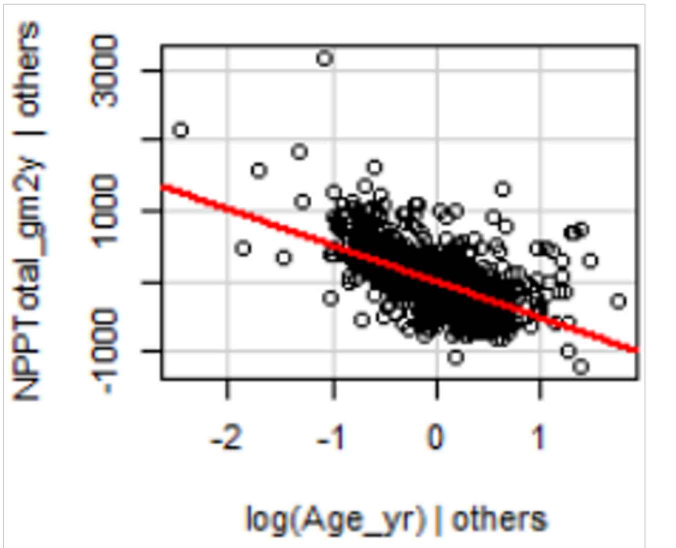
More observations than predictor variables

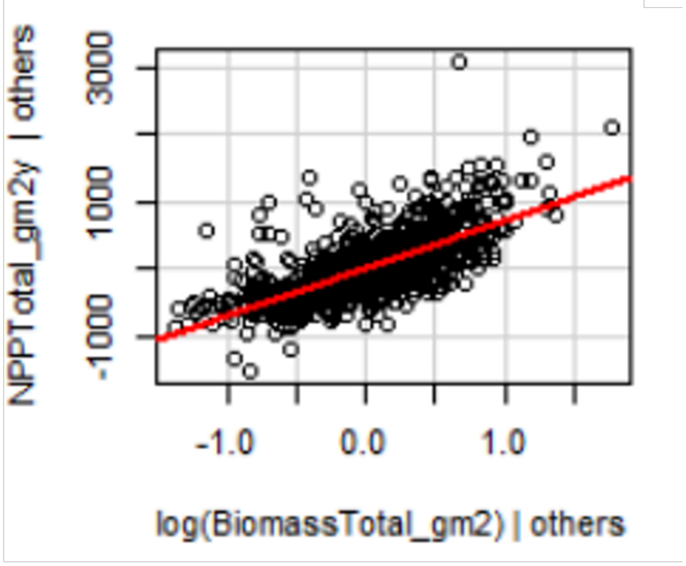
- Ideally at least 10x observations than predictors to avoid "overfitting"
- Uncorrelated predictor variables (assessed using scatterplot matrix; VIFs)
- Linear relationship between Y and each X, holding others constant (non-linearity assessed by AV plots)

Lecture 10: Analyses

Regression of Y vs. each X does not consider effect of other predictors:

want to know shape of relationship while holding other predictors constant





Lecture 10: Collinearity

- Potential predictor variables are often correlated (e.g., morphometrics, nutrients, climatic parameters)
- Multicollinearity (strong correlation between predictors) causes problems for parameter estimates
- Severe collinearity causes unstable parameter estimates: small change in a single value can result in large changes in βp - estimates
- Inflates partial slope error estimates, loss of power

```
> cor(samp[,1:6])
                    Lat        Long NPPTotal_gm2y      Lgs_mo       MAT_C     MAP_mm
Lat           1.0000000 -0.02451430    -0.4002104 -0.60704499 -0.70066118 -0.4463274
Long         -0.0245143  1.00000000    -0.1717559 -0.01783002 -0.08775318 -0.2215868
NPPTotal_gm2y -0.4002104 -0.17175589     1.0000000  0.55946243  0.55177184  0.5278616
Lgs_mo       -0.6070450 -0.01783002     0.5594624  1.00000000  0.92369151  0.5692762
MAT_C        -0.7006612 -0.08775318     0.5517718  0.92369151  1.00000000  0.5788252
MAP_mm       -0.4463274 -0.22158681     0.5278616  0.56927616  0.57882518  1.0000000
```

Lecture 10: Collinearity

Collinearity can be detected by:

- Variance inflation Factors:
  - VIF for Xj=1/ (1-r2)

- VIF > 10 = bad

- Best/simplest solution:
  - exclude variables that are highly correlated with other variables
  - they are probably measuring similar
  - thing and are redundant

Lecture 10: Interactions

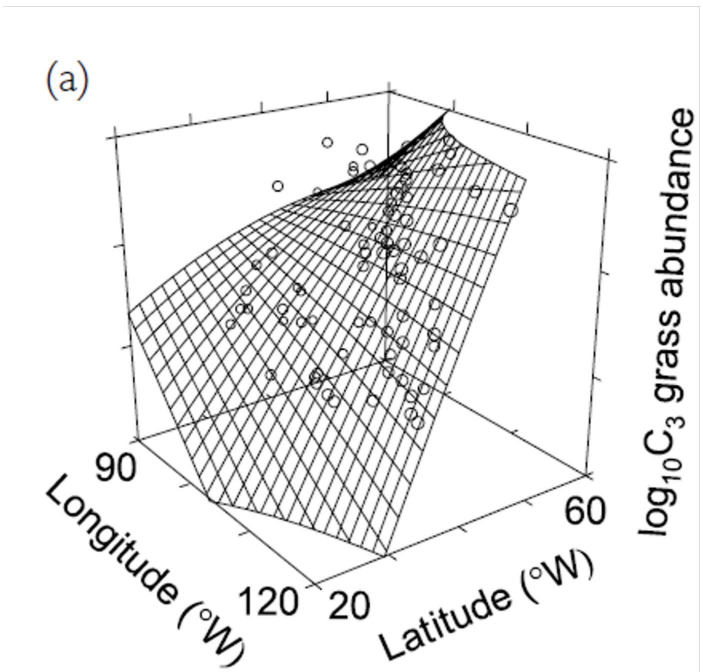
Predictors can be modeled as:

- additive (effect of temp, plus precip, plus fertility) or
- multiplicative (interactive)
- Interaction: effect of Xi depends on levels of Xj
- The partial slope of Y vs. X1 is different for different levels of X2 (and vice versa); measured by β3

$$y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i \quad \text{vs.} \quad y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + +\beta_3 X_{i3} \epsilon_i$$

"Curvature" of the regression (hyper)plane

Lecture 10: Analyses



Lecture 10: Analyses

Adding interactions:

- many more predictors ("parameter proliferation"):
- 2n; 6 params= 64 terms; 7 params= 128
- interpretation more complex
- When to include interactions? When they make biological sense

Lecture 10: Dummy variables

Multiple Linear Regression accommodates continuous and categorical variables (gender, vegetation type, etc.) Categorical vars as "dummy vars", n of dummy variables = n-1 categories

Sex M/F:

- Need 1 dummy var with two values (0, 1)

Fertility L/M/H:

- Need 2 dummy var, each with two values (0, 1): fert1 (0 if L or H, 1 if M), fert2 (1 if H, 0 if L or M)

Fertility	fert1	fert2
Low	0	0
Med	1	0
High	0	1

Lecture 10: Analyses

Coefficients interpreted relative to reference condition

- R codes dummy variables automatically
- picks "reference" level alphabetically
- Dummy variables with more than 2 levels add extra predictor variables to model

Fertility	fert1	fert2
Low	0	0
Med	1	0
High	0	1

Lecture 10: Analyses

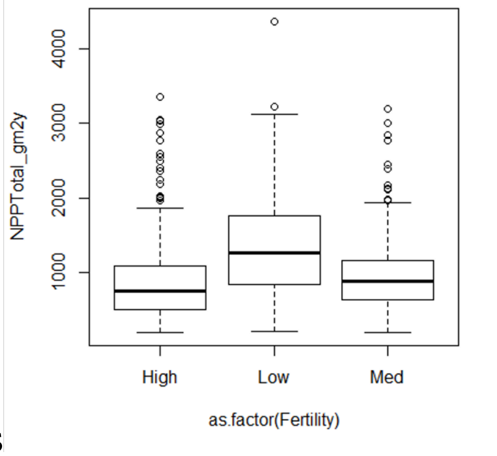
```
> mod1 <- lm(NPPTotal_gm2y ~ MAT_C + MAP_mm + as.factor(Fertility))
> summary(mod1)

Call:
lm(formula = NPPTotal_gm2y ~ MAT_C + MAP_mm + as.factor(Fertility))

Residuals:
     Min       1Q   Median       3Q      Max
-1470.19  -301.59   -49.63   240.82  2901.70

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)             359.81714   31.92698  11.270  < 2e-16 ***
MAT_C                    33.72747    2.73715  12.322  < 2e-16 ***
MAP_mm                    0.34128    0.03443   9.913  < 2e-16 ***
as.factor(Fertility)Low 205.15606   34.36461   5.970  3.1e-09 ***
as.factor(Fertility)Med 116.33231   32.30397   3.601 0.000329 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 456.4 on 1232 degrees of freedom
Multiple R-squared:  0.3899, Adjusted R-squared:  0.3879
F-statistic: 196.8 on 4 and 1232 DF,  p-value: < 2.2e-16
```



S

Lecture 10: Comparing models

When have multiple predictors (and interactions!)

- how to choose "best" model?
- Which predictors to include?
- Occam's razor: "best" model balances complexity with fit to data

To chose:

- compare "nested" models

Overfitting

- getting high r2 just by having more (useless predictors)
- so r2 is not a good way of choosing between nested models

Lecture 10: Comparing models

Need to account for increase in fit with added predictors:

- Adjusted r2
- Akaike's information criterion (AIC)
- Both "penalize" models for extra predictors
- Higher adjusted r2 and lower AIC are better when comparing models

$$\text{Adjusted } r^2 = 1 - \frac{SS_{\text{Residual}}/(n-(p+1))}{SS_{\text{Total}}/(n-1)}$$

$$\text{Akaike Information Criterion (AIC)} = n[\ln(SS_{\text{Residual}})] + 2(p+1) - n\ln(n)$$

Lecture 10: Comparing models

But how to compare models?

- Can fit all possible models
  - compare AICs or adj- r2,
  - tedious w lots of predictors
- Automated forward (and backward) stepwise procedures: start w no terms (all terms), add (remove) terms w largest (smallest)
  - partial F statistic

We will use manual form of backward selection

Lecture 10: Analyses

```
lm(formula = log(NPPTotal_gm2y) ~ log(Age_yr) + log(BiomassTotal_gm2) +
    MAT_C + as.factor(LeafType))

Residuals:
    Min      1Q  Median      3Q     Max
-0.8319 -0.1391 -0.0125  0.1329  1.0132

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                2.106796   0.100738  20.914  < 2e-16 ***
log(Age_yr)               -0.530852   0.015108 -35.138  < 2e-16 ***
log(BiomassTotal_gm2)      0.728076   0.013779  52.838  < 2e-16 ***
MAT_C                      0.007861   0.001320   5.954 3.41e-09 ***
as.factor(LeafType)needle -0.270499   0.014510 -18.642  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2352 on 1215 degrees of freedom      AIC= -62.16
Multiple R-squared:  0.8059,    Adjusted R-squared:  0.8053
```

Lecture 10: Predictors

Usually want to know relative importance of predictors to explaining Y

- Three general approaches:
- Using F-tests (or t-tests) on partial regression slopes
- Using coefficient of partial determination
- Using standardized partial regression slopes

Lecture 10: Predictors

Using F-tests (or t-tests) on partial regression slopes:

- Conduct F tests of Ho that each partial regression slope = 0
- If cannot reject Ho, discard predictor
- Can get additional clues from relative size of F-values
- Does not tell us absolute importance of predictor (usually can not directly compare slope parameters)

Lecture 10: Predictors

Using coefficient of partial determination:

- the reduction in variation of Y due to addition of predictor (Xj)

$$r^2_{X_j} = \frac{SS_{\text{Extra}}}{\text{Reduced } SS_{\text{Residual}}}$$

SSextra

- Increased in SSregression when Xj is added to model

- Reduced SSresidual is the unexplained SS from model without Xj

Lecture 10: Predictors

Using standardized partial regression slopes:

- predictors of predictor variables can not be directly compared
- Why?
- Standardize all vars (mean = 0, sd= 1)
- Scales are identical and larger PRS mean more important variable

Lecture 10: Predictors

Using partial r2 values:

```
Coefficients
                         SSR df pEta-sqr dR-sqr
(Intercept)          24.1975  1   0.2647     NA
log(Age_yr)          68.3055  1   0.5040 0.1972
log(BiomassTotal_gm2) 154.4561  1   0.6968 0.4460
MAT_C                 1.9615  1   0.0284 0.0057
as.factor(LeafType)  19.2262  1   0.2224 0.0555

Sum of squared errors (SSE): 67.2
Sum of squared total  (SST): 346.3
```

Lecture 10: Reporting results

Results are easiest to report in tabular format

```
lm(formula = log(NPPTotal_gm2y) ~ log(Age_yr) + log(BiomassTotal_gm2) +
    MAT_C + as.factor(LeafType))

Residuals:
    Min      1Q  Median      3Q     Max
-0.8319 -0.1391 -0.0125  0.1329  1.0132

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                2.106796   0.100738  20.914  < 2e-16 ***
log(Age_yr)               -0.530852   0.015108 -35.138  < 2e-16 ***
log(BiomassTotal_gm2)      0.728076   0.013779  52.838  < 2e-16 ***
MAT_C                      0.007861   0.001320   5.954 3.41e-09 ***
as.factor(LeafType)needle -0.270499   0.014510 -18.642  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2352 on 1215 degrees of freedom
Multiple R-squared:  0.8059,	Adjusted R-squared:  0.8053
```

```
Coefficients
                          SSR df pEta-sqr dR-sqr
(Intercept)         2885.9581  1   0.7174     NA
log(Age_yr)          435.1076  1   0.2768 0.1228
BiomassTotal_gm2     926.7684  1   0.4491 0.2617
MAT_C                 64.9419  1   0.0540 0.0183
TEB_DD                16.8202  1   0.0146 0.0047
as.factor(LeafType)  268.0125  1   0.1908 0.0757

Sum of squared errors (SSE): 1136.7
Sum of squared total  (SST): 3541.9
```

Lecture 10: Reporting results

Results are easiest to report in tabular format

Response	df	Predictor	coefficient	t	r2	Partial r2	p-value
LN(NPP)	4, 1215	Intercept	2.11	20.9	0.81		<0.0001
		LN(stand age); years	-0.53	-35.1		0.50	<0.0001
		LN(stand biomass); g/m^2	0.73	52.8		0.70	<0.0001
		MAT; °C	0.008	5.95		0.03	<0.0001
		Forest type; needle vs. broadleaf	-0.27	-18.6		0.22	<0.0001